NBER WORKING PAPER SERIES

USING SAMPLES OF UNEQUAL LENGTH IN GENERALIZED METHOD OF MOMENTS ESTIMATION

Anthony W. Lynch
Jessica A. Wachter

Working Paper 14411
http://www.nber.org/papers/w14411

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2008

We thank Yacine Ait-Sahalia, David Chapman, Robert Engle, Martin Lettau, Andrew Lo, Kenneth Singleton, Robert Stambaugh, Jim Stock, Amir Yaron, Motohiro Yogo, as well as seminar participants at the 2005 AFA meetings, at New York University, at the Wharton School and at the University of Pennsylvania Department of Economics for their comments and suggestions. The views expressed herein are those of the author(s) and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2008 by Anthony W. Lynch and Jessica A. Wachter. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Using Samples of Unequal Length in Generalized Method of Moments Estimation
Anthony W. Lynch and Jessica A. Wachter
NBER Working Paper No. 14411
October 2008, Revised June 2011
JEL No. C32,G12

ABSTRACT

Many applications in financial economics use data series with different starting or ending dates. This paper describes estimation methods, based on the generalized method of moments (GMM), which make use of all available data for each moment condition. We introduce two asymptotically equivalent estimators that are consistent, asymptotically normal, and more efficient asymptotically than standard GMM. We apply these methods to estimating predictive regressions in international data and show that the use of the full sample affects point estimates and standard errors for both assets with data available for the full period and assets with data available for a subset of the period. Monte Carlo experiments demonstrate that reductions hold for small-sample standard errors as well as asymptotic ones.

Anthony W. Lynch
New York University
44 W. 4th Street, #9-190
New York, NY 10012
and NBER
alynch@stern.nyu.edu

Jessica A. Wachter
Department of Finance
2300 SH-DH
The Wharton School
University of Pennsylvania
3620 Locust Walk
Philadelphia, PA 19104
and NBER
jwachter@wharton.upenn.edu

Introduction

Many applications in financial economics involve data series that have different starting dates, or, more rarely, different ending dates. Settings where some data series are available over a much shorter time frame than others include estimation and testing using international data, and performance evaluation of mutual funds. These problems represent only the most extreme examples of differences in data length. More broadly, aggregate stock return data may be available over a longer time frame than macroeconomic data, cash flow and earnings data, term structure data, or options data.

The econometrics literature has derived various methods of confronting samples of unequal lengths. An early work is that of Anderson (1957), who derives a maximum likelihood estimator for a bivariate normal distribution in which one variable has more observations than another. More recently, Harvey, Koopman, and Penzer (1998) develop a Kalman-filter approach to missing data, while Schmidt (1977), Swamy and Mehta (1975), and Conniffe (1985) focus on extending the seemingly unrelated regression approach to cases in which more data is available for one equation than the other. Stambaugh (1997) derives estimates of the mean and variance of financial time series, as well as the posterior distribution of returns, assuming returns are normally and independently distributed, in a setting where some return series start at a later date than others. Pastor and Stambaugh (2002a, 2002b) derive Bayesian posteriors for means and variances of mutual fund returns using samples of unequal length, under the assumption of normality and identically and independently distributed returns. Storesletten, Telmer, and Yaron (2004) combine a time series of macroeconomic variables with the shorter Panel Study of Income Dynamics (PSID) to estimate the relationship between cross-sectional variance and recessions. They show how to use the panel structure of the data to infer how the PSID would have behaved over the longer time span. Patton (2006) derives maximum likelihood estimators for samples of unequal length when data are non-normal.[1]

These previous studies take a likelihood-based approach. In contrast, our approach is based on the generalized method of moments (GMM). We show that our method is

[1]See Little and Rubin (2002) for a survey of the statistical literature confronting missing data problems. Another strand of literature considers the problem of n independent individuals observed at up to T time periods, where some individuals drop out of the study (see, e.g., Robins and Rotnitsky (1995)). The independence across individuals and the fact that asymptotics are derived as n, rather than T, approaches infinity differentiates this problem from the one considered here.

more efficient than standard GMM and more efficient than introducing the data from the longer series in a "naive" way. Because it is based on GMM, our method can be used for nonlinear estimation, and for processes that are serially correlated and feature conditional heteroskedasticity. We focus on GMM because, as shown as Cochrane (2001), many common estimation techniques used in finance can be seen as special cases of GMM. Further, many empirical studies in finance explicitly use GMM, such as Harvey (1989), MacKinlay and Richardson (1991) and Zhou (1994).[2] Assumptions required for the consistency and asymptotic normality of the standard GMM estimator are also required here. We adopt the mixing assumption of White and Domowitz (1984) as a means of limiting the temporal dependence of the underlying stochastic process. Intuitively, mixing requires that autocovariances vanish as the lag length increases. This assumption allows for many processes of interest in financial economics, such as finite ARMA processes with general conditions on the underlying errors (see Phillips (1987)).[3]

Because our method is based on GMM, many of our results are asymptotic.[4] So that the asymptotic approximation is reasonable, care must be taken to insure that the missing data problem does not become trivial as the sample size becomes large. We thus develop an asymptotic theory that keeps the fraction of missing data fixed as the sample size approaches infinity. To be precise, if T denotes the length of the longer sample, we say that λT is the length of the shorter sample, for $0 < \lambda \leq 1$. We hold λ constant, as T approaches infinity. This approach has a parallel in the simulated method of moments estimation technique (see Duffie and Singleton (1993)), where the length of the simulated series divided by the length of the observed series is assumed to be constant as both series lengths approach infinity, and also in the literature on structural breaks (Andrews and Fair (1988), Ghysels and Hall (1990), Andrews and Ploberger (1994), Stock (1994), Sowell (1996), Ghysels, Guay, and Hall (1997)).

We focus on the case in which some moment conditions are observed over the full range

[2]Burguete, Gallant, and Souza (1982) and Hansen (1982) describe the GMM estimator and derive its asymptotic properties. Hansen and Singleton (1982) derive implications for estimation and testing of financial models; Brandt (1999) derives implications for the estimation of optimal portfolio and consumption choice. Newey and McFadden (1994) and Hall (2005) survey work on GMM and related estimators.

[3]Like many of the studies mentioned above, we do assume that the data is missing at random, in the sense defined by Little and Rubin (2002). Stambaugh (1997) discusses cases where this assumption holds in financial time series, such as when the start date depends only on the long-history asset returns, and cases where it does not, such as when the decision to add a country to a list of emerging markets depends on past unobserved returns on that country (see Goetzmann and Jorion (1999)).

[4]We also verify, in Monte Carlo experiments, that our methods deliver efficiency gains in small samples.

of dates while others are observed over a time span that has the same ending date but a later starting date because this is the most common pattern in finance applications (we later generalize this to other patterns of missing data).[5] While general, these estimators are straightforward to implement, as we show in an application involving international data (Section 2), and have natural and intuitive interpretations.

The first estimator (which we call the *adjusted-moment* estimator) uses full sample averages to estimate the moments for which full-sample data are available, and short sample averages to estimate moments for which only short-sample data are available. Then the moments for which only the short sample is available are "adjusted" using coefficients from a regression of the short-sample moments on the full-sample moments. This is reminiscent of an adjustment that appears in Stambaugh (1997) and Little and Rubin (2002) but here operates in a more general context. The second estimator, (which we call the *over-identified* estimator) uses the extra data available from the full sample as a new set of moment conditions. This estimator was suggested by Stambaugh (1997) and, in the linear context of that paper, turns out to be identical to our adjusted-moment estimator (and the maximum-likelihood estimator proposed in that paper). In the more general context of our paper, the two estimators are equivalent asymptotically but typically differ in finite samples.

In that it is based on GMM, our study is closely related to that of Singleton (2006, Chapter 4.5). Besides placing the missing data problem within the context of GMM, Singleton also takes the same approach to asymptotics: Namely the ratio of the length of the shorter sample to that of the longer sample remains constant as the total length goes to infinity in both his study and ours. Singleton proposes moment conditions that are the same as those for our over-identified estimator. However, he derives a different weighting matrix. The weighting matrix that we derive allows us to show that our estimators are more efficient than standard GMM, and more efficient than a naive approach to using the full sample. We also depart from Singleton's study in that we define the asymptotically equivalent adjusted-moment estimator, study the finite-sample performance of the estimators, and apply them to predictive regressions.

The organization of the paper is as follows. Section 1 defines our estimators and discusses their efficiency properties. Section 2 provides intuition for the efficiency gains from using the

[5] In its focus on the efficiency results of carefully including additional data, this study has parallels in studies that focus on including high-frequency data in estimation while accounting for market microstructure effects (see Ait-Sahalia, Mykland, and Zhang (2005), Bandi and Russell (2006)).

full sample. Section 3 illustrates our methods through an application to international data. Section 4 presents a Monte Carlo analysis showing that the efficiency gains are present in small samples. Most of our paper focuses on the case where one data series begins earlier than another. However, our approach can easily be generalized to other patterns of missing data, provided that the data is missing in large blocks.[6] Section 5 provides this generalization. Section 6 concludes.

1 GMM estimators for samples of unequal length

1.1 Definitions

Following Cochrane (2001), assume that the model to be estimated can be expressed as

$$E\left[f(x_t, \theta_0)\right] = 0.$$

Here, f is a vector of l restrictions. The true parameters of the model are represented by the q-vector θ_0. Finally, x_t is a vector-valued stochastic process. In what follows, we derive results based on assumptions that are standard in a GMM setting; see Appendix A for more detail.

In many applications, it happens that data are missing for the early part of the sample period for some moment conditions (see Section 2 for an application to international data). We partition the elements of x_t so that $x_t = \left[x_{1t}^\top \; x_{2t}^\top\right]^\top$, where data on x_{1t} are assumed to be available for the full period, and data on x_{2t} are assumed to be available for only the later part of the sample period. Similarly, we can partition the elements of f into those that depend only on x_{1t} and those that depend on both x_{1t} and x_{2t}: $f(x_t, \theta) = [f_1(x_{1t}, \theta)^\top \; f_2(x_t, \theta)^\top]^\top$. Let f_1 be $l_1 \times 1$ and f_2 be $l_2 \times 1$, where $l_1 + l_2 = l$.

Let λ denote the fraction of the period for which all data are available. Then x_{1t} is observed from $t = 1, \ldots, T$, while x_{2t} is observed from $t = (1-\lambda)T + 1, \ldots T$. Define the

[6] Under general assumptions on the dependence of the underlying stochastic process, it is necessary that the number of "blocks" remains fixed asymptotically. This distinguishes the problem we tackle from the problem posed by data sampled at different frequencies (see Ghysels, Santa-Clara, and Valkanov (2005)).

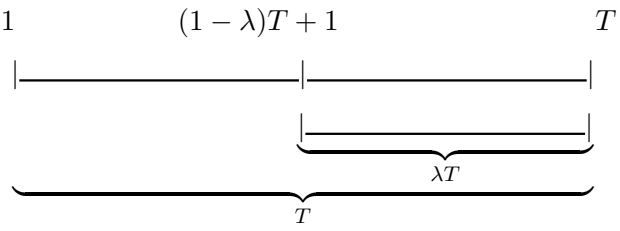

Figure 1: Notation for data missing at the start of the sample

following partial sums:[7]

$$\begin{aligned} g_{1,T}(\theta) &= \frac{1}{T}\sum_{t=1}^{T} f_1(x_{1t},\theta), \\ g_{1,(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f_1(x_{1t},\theta), \\ g_{1,\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_1(x_{1t},\theta), \end{aligned}$$

and

$$g_{2,\lambda T}(\theta) = \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f_2(x_t,\theta).$$

Sums of f are indexed by the length of the sample. This is a slight abuse of notation because the subscript λT does not refer to the sum taken over observations $1,\ldots,\lambda T$. The subscripts λT, $(1-\lambda)T$ and T can be understood as referring to intervals of the data rather than the ending point of the sample. Figure 1 illustrates the notation.

Let $w_{1t} = f_1(x_{1t},\theta_0)$ and $w_{2t} = f_2(x_t,\theta_0)$. Following Hansen (1982), define matrices

$$R_{ij}(\tau) = E\left[w_{i0}w_{j,-\tau}^{\top}\right], \quad i,j = 1,2.$$

Under standard assumptions, these sums converge (see White (1994, Proposition 3.44)). Let

$$S_{ij} = \sum_{\tau=-\infty}^{\infty} R_{ij}(\tau).$$

[7]Formally, λ is a rational number strictly between 0 and 1. Define n_0 to be the smallest positive integer n such that $n\lambda$ is an integer. We consider partial sums of f of length λT and $(1-\lambda)T$ for T a multiple of n_0. For the remainder of the paper, we let T approach infinity along the subsequence of integer multiples of n_0. Alternatively, we could define partial sums of length $\lambda n_0 T'$ and $(1-\lambda)n_0T'$ for any integer T'. The results would be identical, but the notation would be more cumbersome.

and

$$S = \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{22} \end{bmatrix}.$$

It is also useful to define the matrix of coefficients from a regression of the second series on the first:

$$B_{21} = S_{21} S_{11}^{-1}.$$

The residual variance from this regression will be denoted Σ, where

$$\Sigma = S_{22} - S_{21} S_{11}^{-1} S_{12}. \tag{1}$$

Note that S is known as the spectral density matrix.

In this setting, standard GMM corresponds to using moment conditions measured over the subperiod for which all the data are available. That is, the standard GMM estimator solves

$$\min_{\theta} \left[g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top \right]^\top W_T \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}.$$

for a positive definite and symmetric weighting matrix W_T. In what follows, we will focus on the case where the weighting matrix is asymptotically efficient. In the case of standard GMM, this implies that the weighting matrix asymptotically approaches S^{-1}. We let $\hat{S}_T$ denote an estimator of S.[8] Let

$$\hat{\theta}_T^{\mathcal{S}} = \text{argmin}_\theta \left[g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top \right]^\top \hat{S}_T^{-1} \begin{bmatrix} g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}. \tag{2}$$

We call this the *short* estimator.

Standard arguments show that the short estimator is consistent and asymptotically normal. However, the short estimator does not use all of the data available. A natural estimator to consider takes the same form as (2), except $g_{1,\lambda T}(\theta)$ is replaced by its full-sample counterpart, $g_{1,T}(\theta)$. Because this is the simplest estimator that makes use of all of the data, we call this the *long* estimator and let

$$\hat{\theta}_T^{\mathcal{L}} = \text{argmin}_\theta \left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top \right]^\top \left(\hat{S}_T^{\mathcal{L}} \right)^{-1} \begin{bmatrix} g_{1,T}(\theta) \\ g_{2,\lambda T}(\theta) \end{bmatrix}, \tag{3}$$

where $\hat{S}_T^{\mathcal{L}}$ is an estimate of $S^{\mathcal{L}}$, the asymptotic variance of $\sqrt{\lambda T} \left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top \right]^\top$.

[8] Stated more precisely, we choose $\hat{S}_T$ to converge to S almost surely. Convergence for estimates of variance-covariance matrices that follow should be interpreted similarly.

We will argue, however, that the long estimator introduces new data in a suboptimal way. We define two alternative estimators. The first takes $\hat{\theta}_T^{\mathcal{L}}$ as a starting point and adjusts the second set of moment conditions based on sample properties of the first set of moment conditions. To define this estimator, let $\hat{B}_{21,\lambda T}$ be a matrix converging to B_{21}. The *adjusted moment* estimator, $\hat{\theta}_T^{\mathcal{A}}$, solves

$$\hat{\theta}_T^{\mathcal{A}} = \text{argmin}_\theta \left[g_{1,T}(\theta)^\top \; g_{2,T}^{\mathcal{A}}(\theta)^\top \right]^\top \left(\hat{S}_T^{\mathcal{A}} \right)^{-1} \left[\begin{array}{c} g_{1,T}(\theta) \\ g_{2,T}^{\mathcal{A}}(\theta) \end{array} \right], \tag{4}$$

where

$$g_{2,T}^{\mathcal{A}}(\theta) = g_{2,\lambda T}(\theta) + \hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta))$$

and $\hat{S}_T^{\mathcal{A}}$ is an estimate of $S^{\mathcal{A}}$, the asymptotic variance of $\sqrt{\lambda T}\left[g_{1,T}(\theta)^\top \; g_{2,T}^{\mathcal{A}}(\theta)^\top \right]^\top$.

The difference between (3) and (4) lies in the second set of moment conditions, for which only the short sample is available. Because

$$g_{1,T} = (1-\lambda) g_{1,(1-\lambda)T} + \lambda g_{1,\lambda T},$$

the second set of moment conditions for the adjusted-moment estimator, $g_{2,T}^{\mathcal{A}}(\theta)$, can be written as

$$g_{2,T}^{\mathcal{A}}(\theta) = g_{2,\lambda T} + \hat{B}_{21,\lambda T}(g_{1,T} - g_{1,\lambda T}).$$

The expression above illustrates the role of the longer sample in helping to estimate the second set of moment conditions. Consider for example the case where g_1 and g_2 are univariate. If g_1 is below average in the second part of the sample, and if g_1 and g_2 are positively correlated, g_2 is also likely to be below average. Thus the estimate of $E[f_2(x_0,\theta)]$ should be adjusted upward relative to g_2.

Finally, we define an estimator that makes use of longer data sample to add over-identifying restrictions. The *over-identified* estimator solves

$$\hat{\theta}_T^{\mathcal{I}} = \text{argmin}_\theta \left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top g_{2,\lambda T}(\theta)^\top \right]^\top \left(\hat{S}_T^{\mathcal{I}} \right)^{-1} \left[\begin{array}{c} g_{1,(1-\lambda)T}(\theta) \\ g_{1,\lambda T}(\theta) \\ g_{2,\lambda T}(\theta) \end{array} \right], \tag{5}$$

where $\hat{S}_T^{\mathcal{I}}$ is an estimate of $S^{\mathcal{I}}$, the asymptotic variance of $\sqrt{\lambda T}\left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top g_{2,\lambda T}(\theta)^\top \right]$.

1.2 Asymptotic distribution

Theorems B.2 and B.3 in Appendix B show that each estimator is consistent for θ_0 and is asymptotically normal. Standard errors can be obtained using the same results as in

previous work on GMM. Standard errors depend on the derivative of the moment conditions. Define

$$D_{0,i} = E\left[\left(\partial f_i/\partial\theta\right)|_{\theta_0}\right],$$

for $i = 1, 2$, and

$$D_0 = \left[\begin{array}{c} D_{0,1} \\ D_{0,2} \end{array}\right].$$

For the short, long, and adjusted-moment estimators, D_0 is the derivative of the moment condition evaluated at θ_0. As shown in Theorem B.3, the asymptotic distributions of the estimators are normal and centered around θ_0. For example, for the adjusted-moment estimator:

$$\sqrt{\lambda T}(\hat{\theta}_T^{\mathcal{A}} - \theta_0) \to_{\mathrm{d}} N\left(0, \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1}\right) \tag{6}$$

Analogous equations hold for the short and long estimator, where $S^{\mathcal{A}}$ is replaced by S and $S^{\mathcal{L}}$ respectively (recall that the short estimator is standard GMM). Similarly, for the over-identified estimator, the derivative D_0 is replaced by $[D_{0,1}^\top\ D_0^\top]$ and $S^{\mathcal{A}}$ is replaced by $S^{\mathcal{I}}$. In each case, the difference between the estimator and θ_0 is scaled by $\sqrt{\lambda T}$. This an arbitrary choice: we could have equally well have chosen $\sqrt{T}$ (or indeed any constant multiplied by $\sqrt{T}$), and adjusted the variance-covariance matrix in (6) appropriately. Regardless of this choice, it is convenient to keep it the same for all four estimators.

An important practical step in implementing these estimators is obtaining estimates of the spectral density matrices $S^{\mathcal{L}}$, $S^{\mathcal{A}}$, and $S^{\mathcal{I}}$ to substitute into the equations above. Conveniently, these estimates can be obtained with no more difficulty than estimating the matrix S because these matrices can be completely characterized in terms of the submatrices S_{ij} of S. As shown in Theorem B.1:[9]

$$S^{\mathcal{L}} = \left[\begin{array}{cc} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} \end{array}\right] \tag{7}$$

$$S^{\mathcal{A}} = \left[\begin{array}{cc} \lambda S_{11} & \lambda S_{12} \\ \lambda S_{21} & S_{22} - (1-\lambda) S_{21} S_{11}^{-1} S_{12} \end{array}\right] \tag{8}$$

$$S^{\mathcal{I}} = \left[\begin{array}{ccc} \frac{\lambda}{1-\lambda} S_{11} & 0 & 0 \\ 0 & S_{11} & S_{12} \\ 0 & S_{21} & S_{22} \end{array}\right]. \tag{9}$$

[9] Our proposed weighting matrix for the over-identified estimator can be contrasted with that proposed by Singleton (2006). The weighting matrix he proposes is equivalent to the inverse of the matrix given in (9), without the $\lambda/(1-\lambda)$ term in the upper left block.

These formulas show that it suffices to have an estimate of the original spectral density matrix S (see Cochrane (2001) for a discussion). Given such an estimate, $S^{\mathcal{A}}$ and $S^{\mathcal{I}}$ are easily constructed by extracting submatrices: S_{11} is the upper left $l_1 \times l_1$ submatrix, S_{22} is the $l_2 \times l_2$ lower right submatrix, and so on. Such an estimate will generally make use of the last λT observations, because it is necessary to have all series available. In Appendix D, we discuss a means of constructing an estimate of S that uses all of the data.

Underlying these results is the asymptotic independence between non-overlapping samples. That is, $\sqrt{\lambda T} g_{1,(1-\lambda)T}$ and $\sqrt{\lambda T} g_{i,\lambda T}$ are jointly normally distributed and have zero covariance in the limit as the sample size approaches infinity. Please see Appendix A for a formal statement and proof. Asymptotic independence is intuitive: as more and more data become available, the parts of the non-overlapping samples that are close to one another become an ever smaller part of the whole. The samples come to be dominated by terms that are far away and thus nearly independent.

1.3 Efficiency properties

We now compare the asymptotic efficiency of the four estimators. The proof of the following theorem can be found in Appendix C.

Theorem 1. *Assume the short, long, adjusted-moment and over-identified estimators are defined as (2)–(5). Then*

1. *The asymptotic distribution of the adjusted-moment estimator is identical to that of the over-identified estimator.*

2. *The adjusted-moment estimator and over-identified estimator are more efficient than the short estimator.*

3. *The adjusted-moment estimator and over-identified estimator are more efficient than the long estimator.*

Theorem 1 shows that asymptotically, the adjusted-moment and over-identified estimators are the same despite the fact that they take very different forms. The second statement shows that there is indeed an efficiency gain from using the longer sample. Moreover, it is more efficient to use the adjusted-moment or over-identified estimators than to use the longer sample in a "naive" way, as the third statement shows.

In contrast, the long estimator, despite its use of all the data, may not be more efficient than the short estimator. Statement 2 relies on the fact that

$$S - S^{\mathcal{A}} = (1-\lambda) \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{21}S_{11}^{-1}S_{12} \end{bmatrix},$$

is positive semi-definite (that is, S is at least as large, in a matrix sense as $S^{\mathcal{A}}$). However, the analogous quantity for the long estimator,

$$S - S^{\mathcal{L}} = (1-\lambda) \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & 0 \end{bmatrix}$$

will generally not be positive semi-definite. Thus it is not sufficient to simply include the full data in the estimation. The non-overlapping part of the sample must be introduced in precisely the right way to produce a gain in efficiency. The difference between the efficient estimators (the adjusted-moment and over-identified estimators) and the long estimator is especially surprising given that, when attention is restricted to estimating $f_1(x,\theta)$, the three estimators are asymptotically identical. In fact, the gains in efficiency occur because the method uncovers the deviation of $g_{1,\lambda T}$ from zero. Because of the correlation between $g_{1,\lambda T}$ and $g_{2,\lambda T}$, the deviation of $g_{1,\lambda T}$ from zero implies that $g_{2,\lambda T}$ is also likely to deviate from zero. The efficient estimators make use of this information to construct an estimator of the mean of $f_2(x,\theta)$ that improves on $g_{2,\lambda T}$.

The previous results address the case when the efficient weighting matrix for each estimator is used. Sometimes it is of interest to use a weighting matrix that is asymptotically inefficient because of small-sample considerations. As the next theorem shows, there is an efficiency gain for using the full sample in this setting as well. The proof is in Appendix C

Theorem 2. *Assume that the weighting matrices approach a positive-definite matrix W. The adjusted-moment estimator is more efficient than the short estimator and the long estimator.*

1.4 Comparing the efficient estimators

Because the adjusted-moment estimator and the over-identified estimator are asymptotically identical, we refer to them as the *efficient estimators*. The above results raise the question of whether these estimators are identical in finite samples, and, if not, what the differences are. We answer these questions by deriving the first-order conditions that determine the

estimators. For the purpose of this discussion, we assume that $\hat{S}_T^{\mathcal{I}} = S^{\mathcal{I}}$, $\hat{S}_T^{\mathcal{A}} = S^{\mathcal{A}}$ and $\hat{B}_{21,\lambda T} = B_{21}$. However, the results apply as long as these matrices are constructed using the same estimated submatrices of S.

As shown in Appendix C, the first-order condition determining the over-identified estimator is equal to

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0. \quad (10)$$

The first order condition determining the adjusted-moment estimator is

$$\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0. \quad (11)$$

According to Theorem 1, these two first order conditions must be equivalent as $T \to \infty$. Indeed they are, because

$$\lim_{T\to\infty} \left.\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{I}}} = \lim_{T\to\infty} \left.\frac{\partial g_{1,\lambda T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{I}}} = \lim_{T\to\infty} \left.\frac{\partial g_{1,T}}{\partial \theta}\right|_{\hat{\theta}_T^{\mathcal{A}}} = D_{0,1},$$

and

$$\begin{aligned} \frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} D_{0,1} + g_{1,\lambda T}^{\top} S_{11}^{-1} D_{0,1} &= \frac{1}{\lambda}\left((1-\lambda) g_{1,(1-\lambda)T}^{\top} + \lambda g_{1,\lambda T}^{\top}\right) S_{11}^{-1} D_{0,1} \\ &= \frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} D_{0,1}. \end{aligned}$$

For finite T, however, (10) and (11) will generally not be equivalent. Therefore the values of the adjusted-moment and over-identified estimators will differ as well.

There is a special case when the two estimators will be the same, even in finite samples. The estimators will be identical when

$$\frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} = \frac{\partial g_{1,\lambda T}}{\partial \theta},$$

which occurs, for example, when the parameter to be estimated is the mean of x.

1.5 The effect of the full sample

How does including the full sample influence the parameter estimates? For convenience, we consider an often-encountered special case. We assume that the system is exact identified, and, moreover, the first set of moment conditions (of length T) is sufficient to identify a

subset θ_1 of the parameters. That is, θ_1 is exactly identified by those moment conditions available over the full sample. We will call the remaining parameters θ_2.

We first discuss the effect of using the full sample on estimation of θ_1. A natural conjecture is that the long, adjusted-moment, and over-identified estimators all produce the same estimates of θ_1, namely those found by setting g_{1T} equal to zero. This is clearly the case for the adjusted-moment and long estimators. For the over-identified estimator, we use the argument in the previous section to decompose the first-order conditions as follows:

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_1} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_1} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0 \quad (12)$$

and

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_2} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_2} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_2} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0 \quad (13)$$

Under our stated assumptions, f_1 is only a function of θ_1, not of θ_2. Therefore, (13) reduces to

$$\left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta_2} g_{2,\lambda T} = 0$$

Further, because the system is exactly identified, and because f_1 can identify θ_1, it follows that $\frac{\partial}{\partial \theta_2} g_{2,\lambda T}$ is invertible and that

$$g_{2,\lambda T} - B_{21} g_{1,\lambda T} = 0.$$

Therefore,

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta_1} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta_1} = 0 \quad (14)$$

is the first-order condition that identifies θ_1 for the over-identified estimator. As discussed in the section above, this set of equations will in general not be satisfied by the value of θ_1 that sets $g_{1T} = 0$. To summarize, the adjusted-moment estimator gives the same estimate for θ_1 as simply using the full sample. The over-identified estimator gives a possibly different estimate, one that depends on the point in time in which the second series begins. While this dependence is possibly unattractive, (14) nonetheless has an interpretation: it is a weighted average of the moment conditions from the earlier and later parts of the sample,

where the weights are proportional to the derivatives, and thus to the amount of information contained in each part of the sample.

We now ask how including the full sample might effect the standard errors of θ_1 and θ_2. We focus on asymptotic results, so the results for the over-identified and adjusted-moment estimator will be the same.

$$D_0 = \begin{bmatrix} D_{0,1} \\ D_{0,2} \end{bmatrix} = \begin{bmatrix} d_{11} & 0 \\ d_{21} & d_{22} \end{bmatrix},$$

where

$$d_{ij} = \frac{\partial f_i}{\partial \theta_j}, \quad i = 1, 2,$$

and where d_{11} and d_{22} are invertible.

The inverse of D_0 takes the form

$$D_0^{-1} = \begin{bmatrix} d_{11}^{-1} & 0 \\ -d_{22}^{-1} d_{21} d_{11}^{-1} & d_{22}^{-1} \end{bmatrix}.$$

Therefore, the asymptotic variance of the short estimator of θ_1 equals $d_{11}^{-1} S_{11} (d_{11}^{-1})^\top$. Similarly, the asymptotic variance of the efficient estimators of θ_1 (first block of $\left(D_0^\top \left(S^{\mathcal{A}} \right)^{-1} D_0 \right)^{-1}$) can be written as

$$d_{11}^{-1} S_{11}^{\mathcal{A}} (d_{11}^{-1})^\top = \lambda d_{11}^{-1} S_{11} (d_{11}^{-1})^\top.$$

This shows that asymptotic standard errors for the estimates of θ_1 shrink by a factor of $1 - \sqrt{\lambda}$ when the efficient estimators are used rather than the short estimator. As shown above, the second set of moment conditions f_2 has no effect on the estimation of θ_1 (see also Ahn and Schmidt (1995)).

The standard errors of the efficient estimators for θ_2 are determined by the second diagonal block of $\left(D_0^\top \left(S^{\mathcal{A}} \right)^{-1} D_0 \right)^{-1}$, which reduces to

$$\begin{aligned} \left(D_0^\top \left(S^{\mathcal{A}} \right)^{-1} D_0 \right)_{22}^{-1} = d_{22}^{-1} \left[d_{21} d_{11}^{-1} S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}} \right] \left(S_{11}^{\mathcal{A}} \right)^{-1} \left[d_{21} d_{11}^{-1} S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}} \right]^\top (d_{22}^{-1})^\top \\ + d_{22}^{-1} \left[S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}} \left(S_{11}^{\mathcal{A}} \right)^{-1} S_{12}^{\mathcal{A}} \right] (d_{22}^{-1})^\top. \quad (15) \end{aligned}$$

Thus the variance for the second set of parameters can be decomposed into two parts. The first part represents the effect of the first moment conditions on the second variables. The second part represents the variance due only to the residual variance of the second set of moment conditions: $S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}} \left(S_{11}^{\mathcal{A}} \right)^{-1} S_{12}^{\mathcal{A}}$ is the variance-covariance matrix of the second set

of moment conditions conditional on the first. The second diagonal block of $\left(D_0^\top S^{-1} D_0\right)_{22}^{-1}$, which gives the standard errors for θ_2 under standard GMM, has an analogous decomposition:

$$\left(D_0^\top (S)^{-1} D_0\right)_{22}^{-1} = d_{22}^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right](S_{11})^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]^\top (d_{22}^{-1})^\top + d_{22}^{-1}\left[S_{22} - S_{21}(S_{11})^{-1}S_{12}\right](d_{22}^{-1})^\top. \quad (16)$$

Comparing (15) with (16) reveals the source of the efficiency gain. The first term in (15) is equal to λ multiplied by the first term in (16):

$$\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]\left(S_{11}^{\mathcal{A}}\right)^{-1}\left[d_{21}d_{11}^{-1}S_{11}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\right]^\top = \lambda\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]S_{11}^{-1}\left[d_{21}d_{11}^{-1}S_{11} - S_{21}\right]^\top.$$

However the second terms are the same, not surprisingly because they represent the variance of the second moment conditions conditional on the value of the first:

$$S_{22}^{\mathcal{A}} - S_{21}^{\mathcal{A}}\left(S_{11}^{\mathcal{A}}\right)^{-1}S_{12}^{\mathcal{A}} = S_{22} - S_{21}S_{11}^{-1}S_{12}.$$

The percent decline in standard errors depends on the first term relative to the whole. For example, when the second set of moments are perfectly correlated with the first set, the residual variance is zero,

$$S_{22} - S_{21}S_{11}^{-1}S_{12} = 0, \quad (17)$$

and the standard errors for θ_2 also shrink by a factor of $1 - \sqrt{\lambda}$. At the other extreme, suppose that f_2 tells you nothing about θ_1, i.e. $d_{21} = 0$ (θ_1 does not enter into f_2) and $S_{21} = S_{12}^\top = 0$ (the moment conditions are independent). Then the inclusion of the longer series leads to no shrinkage in the asymptotic variance of θ_2.

Of course, even if the two sets of moment conditions are independent ($S_{21} = S_{12}^\top = 0$), the sampling variance of θ_2 may still fall because the sampling variance of θ_1 is reduced. As long as $d_{21} \neq 0$, the first term in (15) is nonzero and there is an effect on the standard errors of θ_2. Similarly, even if there is no impact of θ_1 on the second set of moment conditions ($d_{21} = 0$) the first set of moment conditions help to estimate θ_2 if the covariance between the two moment conditions is nonzero.

Imposing the restriction $d_{21} = 0$ allows us to extend the above discussion to the long estimator. In this exactly-identified case, the long estimator $\hat{\theta}_T^{\mathcal{L}}$ solves

$$g_{1,T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0 \quad (18)$$

$$g_{2,\lambda T}(x, \hat{\theta}_T^{\mathcal{L}}) = 0. \quad (19)$$

It follows that long estimates for θ_1 are asymptotically identical to the efficient estimates for these parameters (they are numerically identical to the adjusted-moment estimates and asymptotically identical to the over-identified estimates). However, the long estimates of θ_2 are numerically identical to the short estimates, not to the efficient estimates.[10] This follows because the efficient estimates for θ_2 solve

$$g_{2,\lambda T}(x, \hat{\theta}_T^{\mathcal{A}}) + \hat{B}_{21,\lambda T}(g_{1,T}(x, \hat{\theta}_T^{\mathcal{A}}) - g_{1,\lambda T}(x, \hat{\theta}_T^{\mathcal{A}})) = 0$$

rather than (19). This starkly illustrates the surprising role of the long sample in helping to estimate θ_2.[11] As we illustrate in the section that follows, this surprising result occurs because the separation uncovers the deviation of $g_{1,\lambda T}$ from zero. Because of the correlation between $g_{1,\lambda T}$ and $g_{2,\lambda T}$, the deviation of $g_{1,\lambda T}$ from zero implies that $g_{2,\lambda T}$ is also likely to deviate from zero. The efficient estimators make use of this information to construct an estimator of the mean of $f_2(x, \theta)$ that improves on $g_{2,\lambda T}$.

2 Application to predictive regressions in international data

This section applies our method to estimating predictive regressions for returns in international data. Reliable international data typically begin substantially later than U.S. data. At the same time, predictive regressions are often measured with noise, making it desirable to use as long a data series as possible. Our methods allow international data to be used at the same time as longer US data.

2.1 Data

For the U.S., we use the annual data of Shiller (1989, Chap. 26), which begin in 1871 and are updated through 2005. Stock returns, prices and earnings are for the S&P 500 index. The predictor variable we use is the ratio of previous ten-year earnings to current stock price. We refer to this as the smoothed earnings-price ratio. This ratio is motivated by the present-value formula linking the earnings-price ratio to returns; normalizing by smoothed earnings rather than earnings has the advantage that it eliminates short-term cyclical noise

[10] It is tempting to conclude that the lower variance for $\hat{\theta}_{1,T}^{\mathcal{L}}$ and the same variance for $\hat{\theta}_{2,T}^{\mathcal{L}}$ implies that the long estimator is more efficient than the short estimator. This is not the case however. Efficiency requires that any linear combination of $\hat{\theta}_{1,T}^{\mathcal{L}}$ and $\hat{\theta}_{2,T}^{\mathcal{L}}$ have lower variance than the same linear combination of short estimates.

[11] The result is even more surprising given that the presence of the second set of moment conditions does not affect estimation of the first set in this exactly identified case, as shown by Ahn and Schmidt (1995).

(see Campbell and Shiller (1988), Campbell and Thompson (2008)). The riskfree rate is the return on six-month commercial paper purchased in January and rolled over in July. Because the first ten years of the sample are used to construct the predictor variable, the data series of the predictor and returns begins in 1881 and ends in 2005. All variables are deflated using the consumer price index (CPI).

Data on international indices come from Ken French's website. The raw data on international indices come from Morgan Stanley's Capital International Perspectives (MSCI). Fama and French (1989) discuss details of the construction of these data. The EAFE is a value-weighted index for Europe, Australia, and the Far East: within the EAFE, countries are added when data become available. For each country returns are value-weighted and countries are then weighted in proportion to their market values in the index. We also examine results for sub-indices. These are Asia-Pacific (Australia, Hong Kong, Japan, Malaysian, New Zealand, Singapore), Europe without the UK (Austria, Belgium, Switzerland, Germany, Spain, France, Italy, Netherlands), Europe with the UK (same as previous with Great Britain and Ireland) and Scandinavia (Denmark, Finland, Norway, Sweden). Data are monthly from 1975 to 2005. We compound the monthly dollar returns on these indices to create annual returns. We then subtract changes in the CPI from the Shiller data set described above from the log of these returns to create real continuously compounded returns.

2.2 Applying the estimators

Let $r_{1,t}$ denote the excess return on the long-history asset (the S&P 500) and $r_{2,t}$ the excess return on the short-history asset (the EAFE or one of the sub-indices). We estimate the predictive regressions

$$r_{1,t+1} = \alpha_1 + \beta_1 z_t + \epsilon_{1,t+1} \tag{20}$$

$$r_{2,t+1} = \alpha_2 + \beta_2 z_t + \epsilon_{2,t+1} \tag{21}$$

jointly for S&P 500 and international index excess returns, where z_t is the smoothed earnings-price ratio on the S&P. Moment conditions are determined by

$$f_1(x_t, \theta) = \begin{bmatrix} 1 \\ z_t \end{bmatrix} (r_{1,t+1} - \alpha_1 - \beta_1 z_t) \tag{22}$$

$$f_2(x_t, \theta) = \begin{bmatrix} 1 \\ z_t \end{bmatrix} (r_{2,t+1} - \alpha_2 - \beta_2 z_t), \tag{23}$$

where $x_t = (r_{1,t}, r_{2,t}, z_{t-1})$,

$$\theta_i = [\alpha_i, \beta_i]^\top, \quad i = 1, 2,$$

and $\theta = [\theta_1^\top \ \theta_2^\top]^\top$. The regression coefficients are identified by the conditions

$$E\left[f_1(x_t, \theta_0)\right] = E\left[f_2(x_t, \theta_0)\right] = 0.$$

The system is exactly identified and f_1 is sufficient to identify α_1 and β_1. Therefore we are in the setting of Section 1.5. Moreover, α_1 and β_1 do not appear as arguments in f_2. The source of the gain in estimating α_2 and β_2 will therefore be the correlation in the moment conditions, which arises from the correlation between shocks to $r_{1,t}$ and $r_{2,t}$, as shown in Section 1.5. We refer to the two moment conditions implied by f_1 as the long-history moment conditions and the moment conditions implied by f_2 as the short-history moment conditions.

Define matrices

$$Z_T = \begin{bmatrix} 1 & z_0 \\ \vdots & \vdots \\ 1 & z_{T-1} \end{bmatrix}, \ Z_{\lambda T} = \begin{bmatrix} 1 & z_{(1-\lambda)T} \\ \vdots & \vdots \\ 1 & z_{T-1} \end{bmatrix}, \ Z_{(1-\lambda)T} = \begin{bmatrix} 1 & z_0 \\ \vdots & \vdots \\ 1 & z_{(1-\lambda)T-1} \end{bmatrix}.$$

and similarly,

$$R_{1,T} = \begin{bmatrix} 1 & r_{1,1} \\ \vdots & \vdots \\ 1 & r_{1,T} \end{bmatrix}, \ R_{1,\lambda T} = \begin{bmatrix} 1 & r_{1,(1-\lambda)T+1} \\ \vdots & \vdots \\ 1 & r_{1,T} \end{bmatrix}, \ R_{1,(1-\lambda)T} = \begin{bmatrix} 1 & r_{1,1} \\ \vdots & \vdots \\ 1 & r_{1,(1-\lambda)T} \end{bmatrix},$$

and

$$R_{2,\lambda T} = \begin{bmatrix} 1 & r_{2,(1-\lambda)T+1} \\ \vdots & \vdots \\ 1 & r_{2,T} \end{bmatrix}.$$

The partial sums in Section 1.1 can then be written as

$$g_{1,T}(x, \theta) = \frac{1}{T} Z_T^\top \left(R_{1,T} - Z_T \theta_1\right) \tag{24}$$

$$g_{1,(1-\lambda)T}(x, \theta) = \frac{1}{(1-\lambda)T} Z_{(1-\lambda)T}^\top \left(R_{1,(1-\lambda)T} - Z_{(1-\lambda)T} \theta_1\right) \tag{25}$$

$$g_{1,\lambda T}(x, \theta) = \frac{1}{\lambda T} Z_{\lambda T}^\top \left(R_{1,\lambda T} - Z_{\lambda T} \theta_1\right) \tag{26}$$

$$g_{2,\lambda T}(x, \theta) = \frac{1}{\lambda T} Z_{\lambda T}^\top \left(R_{2,\lambda T} - Z_{\lambda T} \theta_2\right) \tag{27}$$

The short estimator is the solution to equations defined by setting (26) and (27) to zero. This is the same as ordinary least squares (OLS) regression over the 1975–2005 period. The adjusted-moment estimator requires an estimate of B_{21}. In this context, this is a 2×2 matrix of coefficients of a multivariate regression of errors from the short-history moment conditions on errors from the long-history moment conditions. To calculate this regression, we first estimate the system using the short method and evaluate f_1 and f_2 at the short estimates. We then have a sequence of observations on the errors for the moment conditions from 1975–2005. Regressing the errors that correspond to f_2 on the errors that correspond to f_1 yields the 4 entries of the matrix $\hat{B}_{21,\lambda T}$. Given $\hat{B}_{21,\lambda T}$, the adjusted-moment estimator is the solution to equations defined by setting (24) and

$$g_{2,\lambda T}(x,\theta) + \hat{B}_{21,\lambda T}\left(g_{1,T}(x,\theta) - g_{1,\lambda T}(x,\theta)\right)$$

to zero. For the long-history asset, this corresponds to OLS regression over the 1881–2005 period. For the short-history asset, this corresponds to a regression over the later part of the sample period, plus an adjustment which, as we show below, can be quite substantial.

While the adjusted-moment and short estimators are exactly identified, the over-identified estimator is not, as its name suggests. Moment conditions for the over-identified estimator are (25), (26) and (27). The weighting matrix is the inverse of an estimate of $S^{\mathcal{I}}$, which can be calculated based on submatrices of an estimate of S as in (9). Below, we explain how we estimate S.

Obtaining standard errors requires an estimate of the derivative matrix D_0 and an estimate for the variance matrix S. The results of Section 1 require only that we choose estimators that are consistent. However, it is most in the spirit of our approach to use the full data in constructing $\hat{D}_T$ and $\hat{S}_T$. A consistent estimator of the derivative matrix D_0 that makes use of the full sample is

$$\hat{D}_T = I_2 \otimes \frac{1}{T} Z_T^\top Z_T,$$

where I_2 is the 2×2 identity matrix.

To construct an estimate for $\hat{S}_T$ that makes use of the full sample, we apply the procedure outlined in Stambaugh (1997) for constructing a positive-definite variance-covariance matrix for data of unequal lengths. We describe this procedure in Appendix D.

2.3 Results

Prior to reporting the results for the predictive regressions, we briefly discuss the estimates of the mean returns implied by our methods. The implementation for this estimation is very similar to, and is less complicated than, the implementation of the predictability estimation described above. Note that our estimators take the same form as those of Stambaugh (1997) in the setting of estimating sample means.[12]

The first two columns of Table 1 report results and standard errors for the short estimator; the second two columns report results and standard errors for the adjusted-moment and over-identified estimators. Because these are numerically equivalent in the setting of estimating means, we refer to them jointly as the efficient estimator. As the columns for short show, the sample mean for excess returns on the S&P 500 in the 1975–2005 period was 5.64% with a standard error of 3.16%. The sample mean over the full period is 3.96% as the efficient column shows. It is also estimated much more precisely: the standard error falls from 3.16% to 1.55%.

Introducing data from 1881–1975 also results in more precise estimates of the excess return on short-history assets. For the EAFE index, the standard errors falls from 3.79 for the short method to 3.09 for the efficient methods (the correlation between the S&P 500 and the EAFE portfolios is 0.67). It is this correlation that leads to the reduced standard errors. In particular, the fact that the mean return for the S&P 500 was somewhat higher in the later part of the sample than the earlier part implies that shocks during the 1975–2005 period had a positive mean on average. The efficients estimators therefore adjust the mean excess return on the EAFE downward.

Estimation for the sub-indices also improves, more dramatically for the European indices and less so for the Asia-Pacific index. While the correlation between the Asia-Pacific index and the S&P 500 index is 0.43, the correlation between the European indices and the S&P 500 exceed 0.70. As shown in Section 1.5, higher correlations between the moment conditions lead to greater improvement for the short-history asset.

Table 2 reports results of estimating the predictive regressions. We show the coefficients

[12] In the i.i.d. normal setting of Stambaugh (1997), the estimate of the matrix B_{21} is comprised of regression coefficients of the short-history series on the long-history series. This estimate will also be consistent under more general distributional assumptions, including conditional heteroskedasticity. However, allowing for serial correlation would require a different estimate of B_{21} (which could be derived from submatrices of S). For the current application (which uses annual non-overlapping observations), allowing for serial correlation is unlikely to have a large effect.

on the predictive variable, the standard error on this variable, and the R^2, computed as the sample variance of the predicted return divided by the sample variance of the total return.[13] In the short sample, the point estimates for all the portfolios are positive but insignificant: t-statistics are below 1 for all portfolios. The R^2 values are also small, e.g. 0.6% for the S&P 500. There is more evidence for predictability in the longer sample. For the S&P 500, the adjusted-moment method leads to an estimated coefficient of 0.093 with a standard error of 0.038 and an R^2 of 4.2%. The over-identified method leads to an estimated coefficient of 0.065 and an R^2 of 2.1%.

Well-known theoretical results demonstrate that ordinary least squares regression produces the best fitting regression estimates, given a single series of data. In this predictive regression setting, OLS is equivalent to our short method. Our results show that one can improve on OLS if one has data on a series for which a longer sample is available. Indeed, this application shows that including the earlier period of the sample has a substantial impact on the estimation for the EAFE and other short-history assets. The adjusted-moment method leads to an estimated coefficient of 0.128, as opposed to 0.073 using the short method. Moreover, the standard error on this estimate falls from 0.118 to 0.097. The implied R^2 is 12%, up from 3.8% when the short method is used. Results for the over-identified estimator are similar: the coefficient is 0.101 with an R^2 of 7.3%. Similar effects are present for sub-indices of the EAFE.

Essentially, our methods exploit the information that the evidence for predictability in the U.S. is stronger in the full sample than over the latter half. Under the assumption of stationarity, shocks to returns and to the predictive variable over the latter half of the sample must be such that the predictive coefficient estimated over this data range is too small. Because of the correlation between international returns and U.S. returns, OLS regression for international returns over the same data range would also be likely to understate the extent of predictability. Thus the efficient estimators adjust the OLS (short) estimate upward. The resulting estimates have less noise, as represented by the smaller standard errors in Table 2. While the annual data sample is still too short to comment on statistical significance (except for the U.S.), it is clear that the economic significance of predictability is a great deal higher when the early part of the sample is included.

[13] For each series, we use the data that are available in computing the R^2.

2.4 Efficient versus inefficient use of the full data

Finally, we use this application to contrast the efficient estimators with the long estimator, which uses the full sample but in an inefficient way. In so doing, we illustrate the theoretical results presented at the end of Section 1.5.

The results in Section 1.5 imply that the long estimate for the predictability coefficient β_1 is numerically identical to the adjusted-moment estimate and asymptotically equal to the over-identified estimate of this coefficient. Both the long and the adjusted-moment estimate are equal to the value obtained from an OLS regression of S&P 500 returns on the predictor variable over the full sample of data. In contrast, the long estimate for β_2 is not equal to the adjusted-moment or over-identified estimate. Rather it is equal to the short estimate of 0.073 (in the case of the EAFE), which is the value obtained from an OLS regression of EAFE returns on the predictor variable over the 1975–2005 period. The adjusted-moment estimate and the over-identified estimate are substantially higher, at 0.128 and 0.101 respectively.

The efficient estimators differ from the long estimator in that they divide the data on the S&P 500 into two moment conditions, one defined over 1881–1975 and one defined over 1975–2005. Dividing the data in this way does not alter the estimate (asymptotically) of the predictive coefficient for the S&P 500. However, this division does create more information: it uncovers the fact that there is less predictability in the S&P 500 over the 1975–2005 period than over the full period. As discussed in the previous section, our efficient estimators correctly incorporate this information to improve estimation of β_2.

3 Monte Carlo Analysis

In the previous sections we introduced two methods of implementing GMM with unequal sample lengths and showed that these methods lead to improvements in asymptotic efficiency. More precise estimates can be obtained both for assets with data available for the full period and, more surprisingly, for assets with data available for the later part of the period. A natural question is whether these gains are present for the small-sample distribution of the estimates.

In this section we answer this question using a Monte Carlo experiment modeled after the estimation of predictability. It is particularly useful to investigate this case in a small-sample setting, as it is well known that asymptotic properties can fail noticeably for predictive

regressions when the regressors are persistent (e.g., Cavanagh, Elliott, and Stock (1995), Nelson and Kim (1993), Stambaugh (1999)).[14]

We simulate from the system (20)–(21) using the adjusted-moment regression coefficients to determine the data-generating process. We augment this system with an autoregression for the log of the smoothed earnings-price ratio z_t:

$$z_{t+1} = \rho_0 + \rho_1 z_t + \epsilon_{z,t+1}. \tag{28}$$

We assume that the shocks are iid and normally distributed. Estimates for ρ_0 and ρ_1 are obtained using the full data set and are equal to -0.294 and 0.892 respectively. For each index, we estimate the variance-covariance matrix of errors from (20), (21) and (28) using the method described in Appendix D. Table 3 reports the variances and correlations. The contemporaneous correlation between innovations to z_t and to S&P 500 returns $r_{1,t}$ is -0.91: this large negative value is due to the fact that price is in the denominator of the smoothed earnings-price ratio. Innovations to z_t are also negatively correlated with innovations to returns on the short-history assets. For example, the correlation with innovations to returns on the EAFE is -0.515. Innovations to returns on the S&P 500 are also highly correlated with innovations to international returns: this correlation is 0.65 for the EAFE and over 0.70 for the European sub-indices. Therefore it is reasonable to expect that incorporating the earlier data period will affect the precision of the estimates for the short-history assets.

For each international index, we simulate 50,000 samples of returns on the S&P 500, values for the predictor variable, and returns on that index. The sample length for the S&P 500 (the long-history asset) and the predictor variable is 124 years; the sample length for the short-history asset is 30 years. We repeat the short, adjusted-moment and over-identified estimations in each. We report both the standard deviations of the estimates (Table 4), and the bias (Table 5, measured as the difference between the mean estimated and the true coefficient).

Table 4 shows that the asymptotic efficiency gains discussed in Section 2.3 also appear in finite samples. For the long-history asset the standard deviation of the predictive coefficient falls from 0.133 to 0.48 for both the adjusted-moment and over-identified methods. For the short-history assets, there is improvement in all but one case (when this asset is calibrated to the Asia-Pacific index). When the asset is calibrated to the EAFE for example, the short method delivers a standard deviation of 0.156. The adjusted-moment method delivers

[14] In contrast, the small-sample standard errors for the means are nearly identical to the asymptotic ones.

a standard deviation of 0.134, the over-identified method a standard deviation of 0.135. When the asset is calibrated to the European index, the standard deviation of the estimate falls from 0.156 to 0.116 for both the adjusted-moment and over-identified methods. In each case, the improvement in the small-sample standard errors is of the same magnitude as the asymptotic standard errors.

The theory presented in Section 1 is silent on the subject of bias. However, it is of interest to compare the performance of the efficient estimators to the standard estimator in this regard. It is not surprising that the bias is reduced under the adjusted-moment and over-identified estimators for the long-history asset. Because these estimators are consistent, introducing the longer data should result in a lower bias. Indeed, while the bias for long-history asset is 0.120 under the short estimator, it is 0.028 under the adjusted-moment estimator and 0.015 under the over-identified estimator.

More surprising is the reduction in the bias for the short-history assets. When the short-history asset is calibrated to the EAFE, the bias is 0.083 under the short estimator. Under both the adjusted-moment and over-identified estimators, the bias is about equal to zero (it is in fact very slightly negative for the over-identified estimator). Similar results are apparent when the short-history asset is calibrated to the other indices.

While a full investigation is outside the scope of this study, the form of the estimators gives some insight into the source of the bias reduction. Both the adjusted-moment and the over-identified estimator use the fact that the standard GMM estimates for the long-history asset differ between the full sample and the later part of the sample. Because standard GMM is consistent, some of this difference arises from the bias in the coefficient (because the bias, on average, will be worse in the later part of the sample than in the full sample). Given that the moment conditions are correlated, the bias in estimates for the long-history asset (measured over the later part of the sample) is also likely to appear for the short-history asset (measured over the same period). The estimators can then use the information on the bias for the long-history asset to correct the bias in the short-history asset.

4 Extensions

In this section we briefly outline how our estimators can be extended to more general patterns of missing data. We focus on the over-identified estimator which has a direct

extension.[15]

Consider intervals of the data defined by points in time where at least one sample moment starts or ends. Say these points in time divide the sample up into disjoint intervals $1, \ldots, n$. Let λ_1 denote the ratio of the length of the first region to the length of the entire sample, λ_2 the ratio of the length of the second region to the length of the entire sample, etc. Note that $\sum_{i=1}^{n} \lambda_i = 1$. Define points $t_1, \ldots t_n$ so that the first data segment begins at $t_1 + 1$, the second data segment at $t_2 + 1$, etc. Then

$$g_{\lambda_j T}(\theta) = \frac{1}{\lambda_j T} \sum_{t=t_j+1}^{t_j+\lambda_j T} f(x_t, \theta), \quad j = 1, \ldots, n.$$

For the case described in Section 1, the first segment consist of points 1 to $(1-\lambda)T$, while the second segment consists of points $(1-\lambda)T + 1$ to T. We adopt the same notational convention as in Section 1: $\lambda_j t$ will refer to the length of the segment between $t_j + 1$ and $t_j + \lambda_j T$, and the segment itself.

Let ϕ_i denote the set of moment conditions that are observed in data segment λ_i, and let π_i denote the number of such moment conditions. Define

$$f_{\phi_j}(x_t, \theta) \quad = \quad \Big(f_{i_1}(x_t, \theta), \ \ldots, \ f_{i_{\pi_j}}(x_t, \theta) \Big)^\top,$$

where $\{i_1, \ldots, i_{\pi_j}\} \in \phi_j$ and $i_1 < \cdots < i_{\pi_j}$. Then f_{ϕ_j} are the components of f observed over the segment $\lambda_j T$. Define the $\pi_j \times 1$ vector

$$g_{\phi_j, \lambda_j T}(\theta) = \frac{1}{\lambda_j T} \sum_{t=t_j+1}^{t_j+\lambda_j T} f_{\phi_j}(x_t, \theta)$$

and the $\pi_j \times \pi_j$ matrices

$$R_{\phi_j}(\tau) = E\left[f_{\phi_j}(x_0, \theta_0) f_{\phi_j}(x_{-\tau}, \theta_0)^\top \right]$$

and

$$S_{\phi_j} = \sum_{\tau=-\infty}^{\infty} R_{\phi_j}(\tau).$$

Define

$$h_T^{\mathcal{I}_n}(\theta) = \Big[g_{\phi_1, \lambda_1 T}(\theta)^\top, \ g_{\phi_2, \lambda_2 T}(\theta)^\top, \ldots, \ g_{\phi_n, \lambda_n T}(\theta)^\top \Big]^\top. \tag{29}$$

[15] The extension for the adjusted-moment estimator as well as examples for various patterns of missing data can be found in the working paper Lynch and Wachter (2004).

The $\mathcal{I}_n$ superscript refers to the fact that these are moment conditions for the over-identified estimator, and that there are n non-overlapping intervals. The T subscript refers to the fact that the data length is T.[16] Let $S^{\mathcal{I}_n}$ be the variance-covariance matrix and $\hat{S}_T^{\mathcal{I}_n}$ be an estimate of $S^{\mathcal{I}_n}$. We can then define the extended over-identified estimator as

$$\hat{\theta}_T^{\mathcal{I}_n} = \operatorname{argmin} h_T^{\mathcal{I}_n}(\theta)^\top \left(\hat{S}_T^{\mathcal{I}_n}\right)^{-1} h_T^{\mathcal{I}_n}(\theta). \tag{30}$$

The same consistency and asymptotic normality results go through for the extended over-identified estimator as for the original over-identified estimator. Moreover,

$$S^{\mathcal{I}_n} = \begin{bmatrix} \frac{1}{\lambda_1} S_{\phi_1} & 0 & \dots & 0 \\ 0 & \frac{1}{\lambda_2} S_{\phi_2} & \dots & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \frac{1}{\lambda_n} S_{\phi_n} \end{bmatrix}$$

We now state a result analogous to Theorem 1. That theorem showed that including the data segment for which some data were missing improved efficiency relative to standard GMM. Here we show that including a new data segment improves efficiency relative to the estimator that includes all data but this segment. Without loss of generality, we consider the full over-identified estimator relative to the over-identified estimator defined over the first $n-1$ blocks of data. The proof (available from the authors) is similar to that of Theorem 1.

Theorem 3. *Assume the over-identified estimator $\hat{\theta}_T^{\mathcal{I}_n}$ is defined as (30). Then this estimator is asymptotically more efficient than $\hat{\theta}_{(1-\lambda_n)T}^{\mathcal{I}_{n-1}}$, the analogous estimator that is defined over the first $n-1$ blocks of data.*

5 Conclusion

This paper has introduced two estimators that extend the generalized method of moments of Hansen (1982) to cases where moment conditions are observed over different sample periods. Most estimation procedures, when confronted with data series that are of unequal length, require the researcher to truncate the data so that all series are observed over the same interval. This paper has provided an alternative that allows the researcher to use all the data available for each moment condition.

[16] This notation does not, of course, completely define the over-identified estimator. For that, one would need the points at which the data intervals begin, $t_1, \dots, t_n$. These points in turn depend on $\lambda_1, \dots, \lambda_n$ and T.

Under assumptions of mixing and stationarity, we demonstrated consistency, asymptotic normality, and efficiency over both standard GMM and an extension of GMM that uses the full data in a naive way. Our base case assumed that the two series had the same end date but different start dates. We then generalized our results to cases where the start date and the end date may differ over multiple series. In all cases, using all the data produces more efficient estimates. Moreover, the impact of including the non-overlapping portion of the data is not limited to estimating moment conditions which are available for the full period. As long as there is some interaction between the moment conditions observed over the full period and those observed over the shorter period there will be an impact on all the parameters. This interaction can be through covariances between the moment conditions, or through the fact that some parameters appear in both the moment conditions available over the full sample and those available over the shorter sample. In an application of our methods to estimation of conditional and unconditional means in international data, we show that this impact can be large.

Our two estimators are as straightforward to implement as standard GMM and have intuitive interpretations. The adjusted-moment estimator calculates moments using all the data available for each series, and then adjusts the moments available over the shorter series using coefficients from a regression of the short-sample moment conditions on the full-sample moment conditions. The over-identified estimator uses the non-overlapping data to form additional moment conditions. These two estimators are equivalent asymptotically, and superior to standard GMM, but differ in finite samples. We leave the question of which estimator has superior finite-sample properties to future work.

Appendix

A Independence results

Underlying all our results is the asymptotic independence and joint normality of sums taken over disjoint intervals. To achieve this result, we rely on an assumption that is standard in the econometrics literature, namely that the underlying process x_t is mixing. That is, let $\{x_t\}_{t=-\infty}^{\infty}$ denote a p-component stochastic process defined over an underlying probability space $(\Omega, \mathcal{F}, P)$. Let $\mathcal{F}_a^b \equiv \sigma(x_t; a \leq t \leq b)$, the Borel σ-algebra of events generated by $x_a, \ldots, x_b$. Consider a function $f : \mathbf{R}^p \times \Theta \to \mathbf{R}^l$ for Θ, a compact subset of $\mathbf{R}^q$. The function f provides the restrictions that determine θ based on the observations of x_t. Following White and Domowitz (1984), define

$$\alpha(\mathcal{F}, \mathcal{G}) \equiv \sup_{\{F \in \mathcal{F}, G \in \mathcal{G}\}} |P(FG) - P(F)P(G)|$$

for σ-algebras $\mathcal{F}$ and $\mathcal{G}$, and

$$\alpha(s) \equiv \sup_t \alpha\left(\mathcal{F}_{-\infty}^t, \mathcal{F}_{t+s}^{\infty}\right).$$

The process $\{x_t\}$ is said to be α-mixing if $\alpha(s) \to 0$ as $s \to \infty$. This assumption guarantees that autocovariances vanish at arbitrarily long longs. Mixing is a convenient assumption because it allows a trade-off between the speed at which $\alpha(s)$ approaches zero and the conditions required on the function f. An ARMA process, for example, entails relatively fast convergence of $\alpha(s)$, and thus requires only weak conditions on f. For a precise statement of these conditions (which we require to hold for f and its first derivatives), as well as other standard assumptions (namely, stationarity of x_t, uniqueness of θ_0, and θ_0 in the interior of the set on which f is defined) see White and Domowitz.

Define

$$w_t = f(x_t, \theta_0).$$

It is useful to slightly generalize the notation of Section 1. Let

$$\begin{aligned} g_T(\theta) &= \frac{1}{T}\sum_{t=1}^{T} f(x_t,\theta) \\ g_{(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T}\sum_{t=1}^{(1-\lambda)T} f(x_t,\theta) \\ g_{\lambda T}(\theta) &= \frac{1}{\lambda T}\sum_{t=(1-\lambda)T+1}^{T} f(x_t,\theta). \end{aligned}$$

The following lemma states that partial sums taken over disjoint intervals are asymptotically independent.

Lemma A.1. *Let $F \in \mathcal{F}_{-\infty}^{0}$. Let μ be a $1 \times l$ vector, and let c be a scalar. Let*

$$P_g = \lim_{T\to\infty} P\left(\sqrt{T}\mu g_T(\theta_0) < c\right).$$

Then

$$\lim_{T\to\infty} P\left(\left(\sqrt{T}\mu g_T(\theta_0) < c\right) F\right) = P_g P(F).$$

Proof. For any integer T,

$$\sqrt{T} g_T(\theta_0) = \frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t + \frac{1}{\sqrt{T}}\sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} w_t,$$

where $\lfloor\sqrt{T}\rfloor$ is the largest integer less than the square root of T. Then

$$\frac{1}{\sqrt{T}}\sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t = \frac{\lfloor\sqrt{T}\rfloor}{\sqrt{T}}\frac{1}{\lfloor\sqrt{T}\rfloor}\sum_{t=1}^{\lfloor\sqrt{T}\rfloor} w_t \to_{\text{a.s.}} 0$$

as $T \to \infty$, by Theorem 2.3 of White and Domowitz (1984). Because

$$\frac{1}{\sqrt{T}}\sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} w_t \in \mathcal{F}_{\sqrt{T}}^{\infty},$$

$$\left| P\left(\left[\frac{1}{\sqrt{T}}\sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right] F\right) - P\left(\frac{1}{\sqrt{T}}\sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right) P(F)\right| < \alpha(\sqrt{T}).$$

White and Domowitz (1984) show that w_t is α-mixing. Therefore $\alpha(\sqrt{T})$ goes to 0 as $T \to \infty$. By the Slutsky theorem,

$$\begin{aligned}
\lim_{T\to\infty} P\left(\left(\sqrt{T}\mu g_T(\theta_0) < c\right) F\right) &= \lim_{T\to\infty} P\left(\left[\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right] F\right) \\
&= \lim_{T\to\infty} P\left(\frac{1}{\sqrt{T}} \sum_{t=\lfloor\sqrt{T}\rfloor+1}^{T} \mu w_t < c\right) P(F) \\
&= P_g P(F),
\end{aligned}$$

where the second line follows because w_t is α-mixing, and the last line follows from a second application of the Slutsky Theorem. □

Lemma A.2. *As $T \to \infty$,*

$$\sqrt{T}\begin{bmatrix} \sqrt{(1-\lambda)}g_{(1-\lambda)T}(\theta_0) \\ \sqrt{\lambda}g_{\lambda T}(\theta_0) \end{bmatrix} \to_{\mathrm{d}} N\left(0, \begin{bmatrix} S & 0 \\ 0 & S \end{bmatrix}\right).$$

Proof. White and Domowitz (1984, Theorem 2.4) show

$$\sqrt{(1-\lambda)T}g_{(1-\lambda)T}(\theta_0) \to_{\mathrm{d}} N(0, S) \tag{31}$$

and

$$\sqrt{\lambda T}g_{\lambda T}(\theta_0) \to_{\mathrm{d}} N(0, S). \tag{32}$$

Stationarity of x_t implies that random variables $f(x_{-(1-\lambda)T+1}, \theta), \ldots, f(x_{\lambda T}, \theta)$ have the same joint distribution as random variables $f(x_1, \theta), \ldots, f(x_T, \theta)$. Thus partial sums taken over $f(x_{-(1-\lambda)T+1}, \theta), \ldots, f(x_{\lambda T}, \theta)$ have the same distribution as the corresponding partial sums taken over $f(x_1, \theta), \ldots, f(x_T, \theta)$. Define

$$\begin{aligned}
\tilde{g}_{\lambda T}(\theta) &= \frac{1}{\lambda T} \sum_{t=1}^{\lambda T} f(x_t, \theta) \\
\tilde{g}_{(1-\lambda)T}(\theta) &= \frac{1}{(1-\lambda)T} \sum_{t=0}^{(1-\lambda)T-1} f(x_{-t}, \theta).
\end{aligned}$$

It suffices to prove the results for $\tilde{g}_{\lambda T}$ and $\tilde{g}_{(1-\lambda)T}$.

Let $\mathcal{N}(c)$ denote the cumulative distribution function of the standard normal distribution evaluated at c. Let μ_1 and μ_2 be $1 \times l$ vectors such that $\mu_1\mu_1^\top = \mu_2\mu_2^\top = 1$. By Lemma A.1,

$$\lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}\tilde{g}_{(1-\lambda)T}(\theta_0) < c_1, \mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0) < c_2\right) =$$
$$\lim_{T\to\infty} P\left(\mu_1\sqrt{(1-\lambda)T}S^{-1}g_{(1-\lambda)T}(\theta_0) < c_1\right) \lim_{T\to\infty}\left(\mu_2\sqrt{\lambda T}S^{-1}\tilde{g}_{\lambda T}(\theta_0) < c_2\right) = \mathcal{N}(c_1)\mathcal{N}(c_2)$$

for scalars a and b. This shows $\tilde{g}_{\lambda T}(\theta_0)$ and $\tilde{g}_{(1-\lambda)T}(\theta_0)$ are asymptotically independent, and therefore that $g_{\lambda T}(\theta_0)$ and $g_{(1-\lambda)T}(\theta_0)$ are asymptotically independent. The result follows from (31) and (32). □

B Deriving the Asymptotic Distribution

This Appendix derives the asymptotic distribution for the four estimators we consider. For notational convenience, it is useful to define functions h^k that give the moment conditions for each estimator. Besides the assumptions stated in Appendix A, the results in this section also require that the weighting matrices for each estimator converge to a positive definite weighting matrix.

$$\begin{aligned}
h_T^{\mathcal{S}}(\theta) &= \left[g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top \\
h_T^{\mathcal{L}}(\theta) &= \left[g_{1,T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top \\
h_T^{\mathcal{A}}(\theta) &= \left[g_{1,T}(\theta)^\top \; \left(g_{2,\lambda T}(\theta) + \hat{B}_{21,\lambda T}(1-\lambda)(g_{1,(1-\lambda)T}(\theta) - g_{1,\lambda T}(\theta))\right)^\top\right]^\top \\
h_T^{\mathcal{I}}(\theta) &= \left[g_{1,(1-\lambda)T}(\theta)^\top \; g_{1,\lambda T}(\theta)^\top \; g_{2,\lambda T}(\theta)^\top\right]^\top,
\end{aligned}$$

where For $k \in \mathcal{S}, \mathcal{L}, \mathcal{A}, \mathcal{I}$, given a weighting matrix W_T^k, let

$$\theta^k = \text{argmin}_\theta \, h_T^k(\theta)^\top W_T^k h_T^k(\theta),$$

Theorem B.1. *As $T \to \infty$,*

$$\sqrt{\lambda T} h_T^k(\theta_0) \to_{\text{d}} N(0, S^k),$$

where $S^{\mathcal{S}} = S$, and $S^{\mathcal{L}}$, $S^{\mathcal{A}}$ and $S^{\mathcal{I}}$ are defined in (7)–(9).

Proof. The result for the short estimator follows directly from Lemma A.2. To illustrate the proof for the remaining matrices, we derive (8); the proofs of (7) and (9) are similar. In what follows, the argument θ_0 is suppressed and convergence is in the sense of almost surely.

Stationarity implies that $S_{11}^{\mathcal{A}} = \lambda S_{11}$. By Lemma A.2,

$$\begin{aligned}
&\lim_{T\to\infty} E\left[\sqrt{\lambda T}\left(\lambda g_{i,\lambda T} + (1-\lambda) g_{i,(1-\lambda)T}\right)\sqrt{\lambda T}\left(g_{j,(1-\lambda)T} - g_{j,\lambda T}\right)\right)^\top\Big] \\
&\quad = \lim_{T\to\infty}\left(-E\left[\sqrt{\lambda T}\lambda g_{i,\lambda T}\sqrt{\lambda T} g_{j,\lambda T}^\top\right] + E\left[\sqrt{\lambda T}(1-\lambda) g_{i,(1-\lambda T)}\sqrt{\lambda T} g_{j,(1-\lambda)T}^\top\right]\right) \\
&\quad = \lambda S_{ij} - \lambda S_{ij} = 0 \quad (33)
\end{aligned}$$

for $i, j = 1, 2$. Therefore,

$$\begin{aligned} S_{12}^{\mathcal{A}} &= \lim_{T\to\infty} E\left[\sqrt{\lambda T}\left(\lambda g_{1,\lambda T} + (1-\lambda) g_{1,(1-\lambda)T}\right)\sqrt{\lambda T}\left(g_{2,\lambda T} + B_{21}(1-\lambda)(g_{1,(1-\lambda)T} - g_{1,\lambda T})\right)^\top\right] \\ &= \lim_{T\to\infty} E\left[\sqrt{\lambda T}\left(\lambda g_{1,\lambda T} + (1-\lambda) g_{1,(1-\lambda)T}\right)\sqrt{\lambda T} g_{2,\lambda T}^\top\right] \\ &= \lim_{T\to\infty} E\left[\sqrt{\lambda T}\lambda g_{1,\lambda T}\sqrt{\lambda T} g_{2,\lambda T}^\top\right] \\ &= \lambda S_{12}. \end{aligned}$$

The second line follows from (33) and the third and fourth lines follow from Theorem A.2. Using similar reasoning,

$$\begin{aligned} S_{22}^{\mathcal{A}} &= \lim_{T\to\infty} E\left[\sqrt{\lambda T} g_{2,\lambda T}\sqrt{\lambda T} g_{2,\lambda T}^\top\right] - 2\lim_{T\to\infty}(1-\lambda) E\left[\sqrt{\lambda T} g_{2,\lambda T}\sqrt{\lambda T} g_{1,\lambda T}^\top\right] B_{21}^\top \\ &\quad + \lim_{T\to\infty} B_{21}(1-\lambda)^2 E\left[\sqrt{\lambda T}(g_{1,(1-\lambda)T} - g_{1,\lambda T})\sqrt{\lambda T}(g_{1,(1-\lambda)T} - g_{1,\lambda T})^\top\right] B_{21}^\top \\ &= S_{22} - 2(1-\lambda) S_{21} S_{11}^{-1} S_{12} + (1-\lambda)^2\left(\frac{\lambda}{1-\lambda} + 1\right) S_{21} S_{11}^{-1} S_{12} \\ &= S_{22} - (1-\lambda) S_{21} S_{11}^{-1} S_{12}, \end{aligned}$$

which completes the derivation of (8). □

Theorem B.2 establishes consistency of the estimators.

Theorem B.2. *As $T \to \infty$, $\hat{\theta}_T^k \to_{\text{a.s.}} \theta_0$ for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}, \mathcal{I}\}$.*

Proof. White and Domowitz (1984) show that under these assumptions

$$|g_{\lambda T}(\theta) - E f(x_t, \theta)| \to_{\text{a.s.}} 0$$
$$|g_{(1-\lambda)T}(\theta) - E f(x_t, \theta)| \to_{\text{a.s.}} 0$$

as $T \to \infty$ uniformly in $\theta \in \Theta$. By the continuous mapping theorem,

$$h_T^k(\theta)^\top W_T^k h_T^k(\theta) \to_{\text{a.s.}} E[f(x_t, \theta)]^\top W^k E[f(x_t, \theta)]$$

for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\}$, and

$$h_T^{\mathcal{I}}(\theta)^\top W_T^{\mathcal{I}} h_T^{\mathcal{I}}(\theta) \to_{\text{a.s.}} E[f_1(x_{1t}, \theta)^\top \; f(x_t, \theta)^\top]^\top W^{\mathcal{I}} E\begin{bmatrix} f_1(x_{1t}, \theta) \\ f(x_t, \theta) \end{bmatrix}$$

uniformly in θ. The result then follows from Amemiya (1985, Theorem 4.1.1). □

For convenience, define the notation

$$\begin{aligned} D_0^k &= D_0 \quad k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\} \\ D_0^{\mathcal{I}} &= \left[D_{0,1}^\top \; D_{0,1}^\top \; D_{0,2}^\top\right]^\top. \end{aligned}$$

The following theorem establishes asymptotic normality.

Theorem B.3.

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \to_{\mathrm{d}} N\left(0, \left((D_0^k)^\top W^k D_0^k\right)^{-1} \left((D_0^k)^\top W^k S^k W^k D_0^k\right) \left((D_0^k)^\top W^k D_0^k\right)^{-1}\right).$$

Proof. Define

$$D_T^k(\theta) = \frac{\partial h_T^k}{\partial \theta}(\theta)$$

for θ in the interior of Θ. For T sufficiently large, $\hat{\theta}_T^k$ lies in the interior of Θ. By the mean value theorem, there exists a $\tilde{\theta}^k$ in the segment between θ_0 and $\hat{\theta}_T^k$ such that

$$h_T^k(\hat{\theta}_T^k) - h_T^k(\theta_0) = D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0).$$

Pre-multiplying by $D_T^k(\hat{\theta}_T^k)^\top W_T^k$:

$$D_T^k(\hat{\theta}_T^k)^\top W_T^k \left(h_T^k(\hat{\theta}_T^k) - h_T^k(\theta_0)\right) = D_T^k(\hat{\theta}_T^k)^\top W_T^k D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0).$$

By the first-order condition of the optimization problem,

$$D_T^k(\hat{\theta}_T^k)^\top W_T^k D_T^k(\tilde{\theta}^k)(\hat{\theta}_T^k - \theta_0) = -D_T^k(\hat{\theta}_T^k)^\top W_T^k h_T^k(\theta_0).$$

Theorem 2.3 of White and Domowitz (1984) implies that

$$D_T^k(\theta) \to_{\text{a.s.}} E\left[\frac{\partial f}{\partial \theta}(x_t, \theta)\right]$$

for $k \in \{\mathcal{S}, \mathcal{L}, \mathcal{A}\}$, and

$$D_T^{\mathcal{I}}(\theta) \to_{\text{a.s.}} E\left[\begin{array}{c} \frac{\partial f_1}{\partial \theta}(x_{1t}, \theta) \\ \frac{\partial f}{\partial \theta}(x_t, \theta) \end{array}\right]$$

uniformly in θ. Therefore by Theorem B.2 and Amemiya (1985, Theorem 4.1.5),

$$\begin{aligned} D_T^k(\hat{\theta}_T^k) &\to_{\text{a.s.}} D_0^k \\ D_T^k(\tilde{\theta}^k) &\to_{\text{a.s.}} D_0^k. \end{aligned}$$

The result follows from the Slutsky Theorem. □

As in Hansen (1982) choosing the weighting matrix that is a consistent estimator of the inverse variance-covariance matrix is efficient for a given set of moment conditions.

Theorem B.4. *Suppose* $W_{\lambda T}^k \to_{\text{a.s.}} W_k = (S^k)^{-1}$. *Then*

$$\sqrt{\lambda T}(\hat{\theta}_T^k - \theta_0) \to_{\text{d}} N\left(0, \left((D_0^k)^\top \left(S^k\right)^{-1} (D_0^k)\right)^{-1}\right).$$

Moreover, this choice of W^k *is efficient for each estimator.*

C Proofs of the theorems in the text

Proof of Theorem 1

It suffices to compare the asymptotic variances of each estimator because the mean is the same for all of them. That is, it suffices to show that the variance in these expressions is equal for the adjusted-moment and over-identified estimators, and is smaller (in a matrix sense) for these estimators than for the long and short estimator. Equivalently, we show

$$\left([D_{0,1}^\top \; D_0^\top]^\top \left(S^{\mathcal{I}}\right)^{-1} \begin{bmatrix} D_{0,1} \\ D_0 \end{bmatrix}\right)^{-1} = \left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1} \leq \left(D_0^\top S^{-1} D_0\right)^{-1}, \tag{34}$$

and

$$\left(D_0^\top \left(S^{\mathcal{A}}\right)^{-1} D_0\right)^{-1} \leq \left(D_0^\top \left(S^{\mathcal{L}}\right)^{-1} D_0\right)^{-1}. \tag{35}$$

where $A \leq B$ should be interpreted as stating that $B - A$ is positive semi-definite.

We begin by showing the equivalence of the adjusted-moment and over-identified estimators. From (9) and from the expression for the inverse of an invertible matrix it follows that

$$\begin{aligned} \left(S^{\mathcal{I}}\right)^{-1} &= \begin{bmatrix} \frac{1-\lambda}{\lambda} S_{11}^{-1} & 0 \\ 0 & S^{-1} \end{bmatrix} \\ &= \begin{bmatrix} \frac{1-\lambda}{\lambda} S_{11}^{-1} & 0 & 0 \\ 0 & S_{11}^{-1} + B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ 0 & -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix}. \end{aligned}$$

Moreover, it follows from the formula for the matrix inverse (see Green (1997, Chapter 2)) that

$$\left(S^{\mathcal{A}}\right)^{-1} = \begin{bmatrix} \frac{1}{\lambda} S_{11}^{-1} + B_{21}^\top \Sigma^{-1} B_{21} & -B_{21}^\top \Sigma^{-1} \\ -\Sigma^{-1} B_{21} & \Sigma^{-1} \end{bmatrix}.$$

The equality in (34) follows.

To show the remaining statements, we note that it suffices to show $S^{\mathcal{A}} \leq S$, and $S^{\mathcal{A}} \leq S^{\mathcal{L}}$.[17] To show $S^{\mathcal{A}} \leq S$, note that

$$S - S^{\mathcal{A}} = (1-\lambda) \begin{bmatrix} S_{11} & S_{12} \\ S_{21} & S_{21}S_{11}^{-1}S_{12} \end{bmatrix}.$$

For any $l \times 1$ vector $v = [v_1^\top,\ v_2^\top]^\top$,

$$\begin{aligned} v^\top(S - S^{\mathcal{A}})v &= (1-\lambda)\left(v_1^\top S_{11}v_1 + v_1^\top S_{12}v_2 + v_2^\top S_{21}v_1 + v_2^\top S_{21}S_{11}^{-1}S_{12}v_2\right) \\ &= (1-\lambda)(S_{11}v_1 + S_{12}v_2)^\top S_{11}^{-1}(S_{11}v_1 + S_{12}v_2) \geq 0 \end{aligned}$$

because S_{11}^{-1} is positive-semi-definite and $\lambda < 1$. To show that $S^{\mathcal{A}} \leq S^{\mathcal{L}}$

$$S^{\mathcal{L}} - S^{\mathcal{A}} = \begin{bmatrix} 0 & 0 \\ 0 & (1-\lambda)S_{21}S_{11}^{-1}S_{12} \end{bmatrix}$$

which is positive semi-definite by the same reasoning. The first statement of the theorem then implies that $\hat{\theta}_T^{\mathcal{I}}$ is also more efficient than $\hat{\theta}_T^{\mathcal{S}}$ and $\hat{\theta}_T^{\mathcal{L}}$.

Proof of Theorem 2

Define

$$U = WD_0\left(D_0^\top W D_0\right)^{-1}.$$

By Theorem B.3, it suffices to show that $U^\top SU - U^\top S^{\mathcal{A}}U$ and that $U^\top S^{\mathcal{L}}U - U^\top S^{\mathcal{A}}U$ are positive semi-definite. For any vector v,

$$v^\top(U^\top SU - U^\top S^{\mathcal{A}}U)v = (Uv)^\top(S - S^{\mathcal{A}})Uv > 0$$

because $S - S^{\mathcal{A}}$ is positive semi-definite. A similar argument shows that $U^\top S^{\mathcal{L}}U - U^\top S^{\mathcal{A}}U$ is positive semi-definite.

[17] For invertible matrices U_1 and U_2, if $U_1 - U_2$ is positive semi-definite, then $U_2^{-1} - U_1^{-1}$ is positive semi-definite (Goldberger (1964, Chapter 2.7)). It follows that for a conforming matrix M, $(M^\top U_1^{-1}M)^{-1} - (M^\top U_2^{-1}M)^{-1}$ is positive semi-definite.

Derivation of the first-order conditions for the efficient estimators

Differentiating the objective function for the over-identified estimator with respect to θ yields

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} + \begin{bmatrix} g_{1,\lambda T}^{\top} & g_{2,\lambda T}^{\top} \end{bmatrix} \begin{bmatrix} B_{21}^{\top}\Sigma^{-1}B_{21} & -B_{21}^{\top}\Sigma^{-1} \\ -\Sigma^{-1}B_{21} & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \frac{\partial g_{1,\lambda T}}{\partial \theta} \\ \frac{\partial g_{2,\lambda T}}{\partial \theta} \end{bmatrix} = 0,$$

which reduces to

$$\frac{1-\lambda}{\lambda} g_{1,(1-\lambda)T}^{\top} S_{11}^{-1} \frac{\partial g_{1,(1-\lambda)T}}{\partial \theta} + g_{1,\lambda T}^{\top} S_{11}^{-1} \frac{\partial g_{1,\lambda T}}{\partial \theta} + \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(g_{2,\lambda T} - B_{21} g_{1,\lambda T}\right) = 0.$$

Differentiating the objective function for the adjusted-moment estimator with respect to θ yields

$$\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \begin{bmatrix} g_{1,T}^{\top} & \left(g_{2,T}^{\mathcal{A}}\right)^{\top} \end{bmatrix} \begin{bmatrix} B_{21}^{\top}\Sigma^{-1}B_{21} & -B_{21}^{\top}\Sigma^{-1} \\ -\Sigma^{-1}B_{21} & \Sigma^{-1} \end{bmatrix} \begin{bmatrix} \frac{\partial g_{1,T}}{\partial \theta} \\ \frac{\partial g_{2,T}^{\mathcal{A}}}{\partial \theta} \end{bmatrix} = 0,$$

which reduces to

$$\frac{1}{\lambda} g_{1,T}^{\top} S_{11}^{-1} \frac{\partial g_{1,T}}{\partial \theta} + \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T}\right)^{\top} \Sigma^{-1} \frac{\partial}{\partial \theta} \left(B_{21} g_{1,\lambda T} - g_{2,\lambda T}\right) = 0.$$

D Estimating the spectral density matrix using the full data set

Calculating the standard errors requires calculating the spectral density matrix for the adjusted-moment and over-identified estimators. Section 1.2 shows that these matrices can be written in terms of submatrices of S, the spectral density matrix corresponding to the original system of equations. One could estimate S over the short sample using any of the standard estimators, extract the submatrices and construct $S^{\mathcal{A}}$ and $S^{\mathcal{I}}$ accordingly.

However, it is more in the spirit of our approach to estimate S using the full data. Define

$$\hat{w}_{it} = f_i(x_t, \hat{\theta}), \qquad i = 1, 2,$$

where $\hat{\theta}$ could be any consistent estimator of θ. Let

$$\hat{S}_{11,T} = \frac{1}{T} \sum_{t=1}^{T} \hat{w}_{1t} \hat{w}_{1t}^{\top}.$$

Note that $\hat{S}_{11,T}$ is the White (1980) estimator of S_{11}. Define

$$\hat{B}_{21} = \sum_{t=(1-\lambda)T+1}^{T} \hat{w}_{2t}\hat{w}_{1t}^{\top} \left(\sum_{t=(1-\lambda)T+1}^{T} \hat{w}_{1t}\hat{w}_{1t}^{\top} \right)^{-1}.$$

Note that $\hat{B}_{12}$ is the matrix of regression coefficients from a regression of the second set of moment conditions on the first. Finally, define

$$\hat{\Sigma} = \frac{1}{\lambda T} \sum_{t=(1-\lambda)T+1}^{T} (\hat{w}_{2t} - \hat{B}_{21}\hat{w}_{1t})(\hat{w}_{2t} - \hat{B}_{21}\hat{w}_{1t})^{\top}.$$

Then $\hat{\Sigma}$ is an estimator of the residual variance of the regression.

Assuming that the errors are serially uncorrelated (but, allowing for conditional heteroskedasticity),

$$\hat{B}_{12} \rightarrow_{\text{a.s.}} S_{11}^{-1} S_{12}$$

and

$$\hat{\Sigma} \rightarrow_{\text{a.s.}} S_{22} - B_{21} S_{11} B_{21}^{\top}.$$

Therefore

$$\hat{S}_T = \begin{bmatrix} \hat{S}_{11,T} & \hat{S}_{11,T}\hat{B}_{21,\lambda T}^{\top} \\ \hat{B}_{21,\lambda T}\hat{S}_{11,T} & \hat{\Sigma} + \hat{B}_{21,\lambda T}\hat{S}_{11,T}\hat{B}_{21,\lambda T}^{\top} \end{bmatrix} \tag{36}$$

is a consistent estimator of S.

Of course, $\hat{S}_T$ is not the only possible estimator of S that uses all of the data. One could naively use all the data by using the full set of observations to estimate S_{11}, but only the last λT to estimate S_{12} and S_{22}. However, this approach may not produce a positive-definite matrix in finite samples. By contrast (36) is positive definite, as shown by Stambaugh (1997), who makes use of it in a maximum likelihood context. The result that $\hat{S}_T$ is positive definite and consistent does not rely on the assumption of iid normal observations. A related question is whether $\hat{S}_T$ is an efficient estimator of S. Anderson (1957) shows $\hat{S}_T$ is the maximum likelihood estimator of the variance-covariance matrix when errors are normal and iid. We leave the questions of the efficiency (and finite sample) properties of $\hat{S}_T$ in a more general GMM setting to future work.

References

Ahn, Seung C., and Peter Schmidt, 1995, A separability result for GMM estimation with applications to GLS prediction and conditional moment tests, *Econometric Reviews* 14, 19–34.

Ait-Sahalia, Yacine, Per A. Mykland, and Lan Zhang, 2005, How often to sample a continuous-time process in the presence of market microstructure noise, *Review of Financial Studies* 18, 351–416.

Amemiya, Takeshi, 1985, *Advanced Econometrics*. (Harvard University Press Cambridge, MA).

Anderson, T. W., 1957, Maximum likelihood estimates for a multivariate normal distribution when some observations are missing, *Journal of the American Statistical Association* 52, 200–203.

Andrews, Donald W. K., and Ray C. Fair, 1988, Inference in Nonlinear Econometric Models with Structural Change, *Review of Economic Studies* 55, 615–640.

Andrews, Donald W. K., and Werner Ploberger, 1994, Optimal tests when a nuisance parameter is present only under the alternative, *Econometrica* 62, 1383–1414.

Bandi, Frederico M., and Jeffrey R. Russell, 2006, Separating microstructure noise from volatility, *Journal of Financial Economics* 79, 655–692.

Brandt, Michael W., 1999, Estimating portfolio and consumption choice: A conditional Euler equations approach, *Journal of Finance* 54, 1609–1645.

Burguete, Jose F., A. Ronald Gallant, and Geraldo Souza, 1982, On unification of the asymptotic theory of nonlinear econometric models, *Econometric Reviews* 1, 151–190.

Campbell, John Y., and Robert J. Shiller, 1988, Stock prices, earnings, and expected dividends, *Journal of Finance* 43, 661–676.

Campbell, John Y., and Samuel B. Thompson, 2008, Predicting excess stock returns out of sample: Can anything beat the historical average?, *Review of Financial Studies* 21, 1509–1531.

Cavanagh, Christopher L., Graham Elliott, and James H. Stock, 1995, Inference in models with nearly integrated regressors, *Econometric Theory* 11, 1131–1147.

Cochrane, John H., 2001, *Asset Pricing*. (Princeton University Press Princeton, NJ).

Conniffe, Denis, 1985, Estimating regression equations with common explanatory variables but unequal numbers of observations, *Journal of Econometrics* 27, 179–196.

Duffie, Darrell, and Kenneth J. Singleton, 1993, Simulated moments estimation of Markov models of asset prices, *Econometrica* 61, 929–952.

Fama, Eugene F., and Kenneth R. French, 1989, Business conditions and expected returns on stocks and bonds, *Journal of Financial Economics* 25, 23–49.

Ghysels, Eric, Alain Guay, and Alastair Hall, 1997, Predictive tests for structural change with unknown breakpoint, *Journal of Econometrics* 82, 209–233.

Ghysels, Eric, and Alastair Hall, 1990, A test for structural stability of Euler conditions parameters estimated via the generalized method of moments estimator, *International Economic Review* 31, 335–364.

Ghysels, Eric, Pedro Santa-Clara, and Rossen Valkanov, 2005, There is a risk-return trade-off after all, *Journal of Financial Economics* 76, 509–548.

Goetzmann, William N., and Philippe Jorion, 1999, Re-emerging markets, *Journal of Financial and Quantitative Analysis* pp. 1–32.

Goldberger, Arthur S., 1964, *Econometric Theory*. (John Wiley and Sons New York).

Green, William H., 1997, *Econometric Analysis*. (Prentice-Hall, Inc. Upper Saddle River, NJ).

Hall, Alastair R., 2005, *Generalized Method of Moments*. (Oxford University Press Oxford, UK).

Hansen, Lars Peter, 1982, Large sample properties of generalized method of moments estimators, *Econometrica* 50, 1029–1054.

Hansen, Lars Peter, and Ken Singleton, 1982, Generalized instrumental variables estimation of nonlinear rational expectations models, *Econometrica* 50, 1269–1286.

Harvey, Andrew, Siem Jan Koopman, and Jeremy Penzer, 1998, Messy time series: A unified approach, in *Advances in Econometrics, Volume 13* (JAI Press Inc.,).

Harvey, Campbell, 1989, Time-varying conditional covariances in tests of asset pricing models, *Journal of Financial Economics* 24, 289–317.

Little, Roderick J. A., and Donald B. Rubin, 2002, *Statistical analysis with missing data*. (John Wiley & Sons Hoboken, NJ) 2 edn.

Lynch, Anthony W., and Jessica A. Wachter, 2004, Using samples of unequal length in generalized method of moments estimation, New York University Working Paper FIN-05-021.

MacKinlay, A. Craig, and Matthew P. Richardson, 1991, Using Generalized Method of Moments to Test Mean-Variance Efficiency, *The Journal of Finance* 46, pp. 511–527.

Nelson, C. R., and M. J. Kim, 1993, Predictable stock returns: The role of small sample bias, *Journal of Finance* 48, 641–661.

Newey, Whitney K., and Daniel McFadden, 1994, Large sample estimation and hypothesis testing, in R.F. Engle, and D. L. McFadden, eds.: *Handbook of Econometrics, Volume IV* (North-Holland, Amsterdam, The Netherlands).

Pastor, Lubos, and Robert F. Stambaugh, 2002a, Investing in equity mutual funds, *Journal of Financial Economics* 63, 351–380.

Pastor, Lubos, and Robert F. Stambaugh, 2002b, Mutual fund performance and seemingly unrelated assets, *Journal of Financial Economics* 63, 315–349.

Patton, Andrew J., 2006, Estimation of multivariate models for time series of possibly different lengths, *Journal of Applied Econometrics* 21, 147–173.

Phillips, Peter C.B., 1987, Time series regressions with a unit root, *Econometrica* 55, 277–301.

Robins, James M., and Andrea Rotnitsky, 1995, Semiparametric efficiency in multivariate regression models with missing data, *Journal of the American Statistical Association* 90, 122–129.

Schmidt, Peter, 1977, Estimation of seemingly unrelated regressions with unequal numbers of observations, *Journal of Econometrics* 5, 365–377.

Shiller, Robert J., 1989, *Market Volatility*. (MIT Press Cambridge, MA).

Singleton, Kenneth, 2006, *Empirical dynamic asset pricing: Model specification and econometric assessment*. (Princeton University Press Princeton, NJ).

Sowell, Fallaw, 1996, Optimal tests for parameter instability in the generalized method of moments framework, *Econometrica* 64, 1085–1107.

Stambaugh, Robert F., 1997, Analyzing investments whose histories differ in length, *Journal of Financial Economics* 45, 285–331.

Stambaugh, Robert F., 1999, Predictive regressions, *Journal of Financial Economics* 54, 375–421.

Stock, James H., 1994, Unit roots, structural breaks and trends, in R.F. Engle, and D. L. McFadden, eds.: *Handbook of Econometrics, Volume IV* (North-Holland, Amsterdam, The Netherlands).

Storesletten, Kjetil, Chris I. Telmer, and Amir Yaron, 2004, Cyclical dynamics in idiosyncratic labor market risk, *Journal of Political Economy* 112, 695–717.

Swamy, P. A. V. B., and J. S. Mehta, 1975, On Bayesian estimation of seemingly unrelated regressions when some observations are missing, *Journal of Econometrics* 3, 157–169.

White, Halbert, 1980, A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity, *Econometrica* 24, 817–838.

White, Halbert, 1994, *Asymptotic Theory for Econometricians*. (Academic Press, Inc.).

White, Halbert, and Ian Domowitz, 1984, Nonlinear regression with dependent observations, *Econometrica* 52, 143–162.

Zhou, Guofu, 1994, Analytical GMM Tests: Asset Pricing with Time-Varying Risk Premiums, *The Review of Financial Studies* 7, pp. 687–709.

Table 1: Mean Excess Returns on International Indices

Means are estimated for excess returns on international indices. Returns are annual, continuously compounded and in excess of the riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. "Short" denotes estimates obtained using standard GMM; "Efficient" denotes estimates obtained using the adjusted-moment method or over-identified method, which are numerically identical in this application. Results for the "Long" estimator (not shown) are equal to the results for Efficient for the U.S. and the results for Short for all other assets. Standard errors are computed using efficient estimates and are robust to conditional heteroskedasticity. Data for the US span the 1881-2005 period; data for the other indices span the 1975–2005 period. Means and standard errors are reported in percentage terms.

	Short		Efficient	
	Mean	SE	Mean	SE
US	5.64	3.16	3.96	1.55
EAFE	5.29	3.79	3.95	3.09
Asia-Pacific	3.52	4.80	2.43	4.46
Europe	6.46	3.68	4.92	2.68
Europe without UK	5.40	4.17	3.69	3.09
Scandinavia	7.00	4.58	5.27	3.60

Table 2: Predictive Regressions for Excess Returns on International Indices

Predictive regressions are estimated in annual data for excess returns on international indices. The table reports the estimate of the coefficient on the predictor variable (Coef.), the standard error (SE) on this coefficient and the R^2 from the regression. The predictive variable is the log of the smoothed earnings-price ratio. Returns are annual, continuously compounded, and in excess of the riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; "OI" denotes the over-identified method. Results for the "Long" estimator (not shown) are equal to the results for AM for the U.S. and the results for Short for all other assets. Standard errors are computed using AM estimates and are robust to conditional heteroskedasticity. Data for the US span the 1881-2005 period; data for the other indices span the 1975–2005 period.

	Short			AM			OI		
	Coef.	SE	R^2	Coef.	SE	R^2	Coef.	SE	R^2
US	0.036	0.077	0.006	0.093	0.038	0.042	0.065	0.038	0.021
EAFE	0.073	0.118	0.038	0.128	0.097	0.117	0.101	0.097	0.073
Asia-Pacific	0.121	0.185	0.059	0.170	0.175	0.117	0.147	0.175	0.086
Europe	0.038	0.103	0.012	0.097	0.076	0.077	0.068	0.076	0.037
Europe without UK	0.018	0.114	0.002	0.080	0.088	0.040	0.050	0.088	0.016
Scandinavia	0.015	0.164	0.001	0.093	0.139	0.044	0.058	0.139	0.017

Table 3: Monte Carlo Parameters for Predictive Regressions

Standard deviations and correlations are estimated in annual data for errors from predictive regressions for use in constructing simulated data. Right-hand-side variables are the US return, an international index return (EAFE or sub-index of the EAFE) and the predictor variable, the log of the smoothed earnings-price ratio. Returns are annual, continuously compounded, and in excess of the riskfree rate. US refers to returns on the S&P 500; EAFE refers to returns on an index for Europe, Asia and the Far East; Asia-Pacific, Europe, Europe without UK and Scandinavia are sub-indices of the EAFE. Data for the US span the 1881-2005 period; data for the other indices span the 1975–2005 period. Predictive coefficients for returns are reported in Table 2 under the heading "AM". The coefficient for the predictor variable is 0.89. Data on international index returns are annual and span the 1975–2005 period. Data on US returns are annual and span the 1881–2005 period.

	Standard deviation	Correlation with $\log(E/P)$	Correlation with U.S.
$\log(E/P)$	0.179		
US	0.170	-0.912	
EAFE	0.207	-0.515	0.653
Asia-Pacific	0.259	-0.309	0.409
Europe	0.205	-0.616	0.775
Europe without UK	0.229	-0.666	0.769
Scandinavia	0.255	-0.578	0.710

Table 4: Predictive Regressions in Repeated Samples

50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. The table reports standard deviations of estimates of the predictive coefficient. In each set of samples there is a long-history asset calibrated to the S&P 500 and a short-history asset calibrated to the EAFE or sub-index. The long-history asset has 124 years of data; the short-history asset has 30 years of data. Predictive coefficients are reported in Table 2 under the heading "AM" and standard deviations and correlations of errors in Table 3. The predictor variable has an autocorrelation coefficient of 0.89. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; "OI" denotes the over-identified method.

	Short	AM	OI
US	0.133	0.048	0.048
EAFE	0.156	0.134	0.135
Asia-Pacific	0.193	0.196	0.197
Europe	0.156	0.116	0.116
Europe without UK	0.175	0.130	0.131
Scandinavia	0.194	0.156	0.157

Table 5: Bias in Predictive Coefficients

50,000 samples of returns are simulated assuming joint normality of excess returns and the predictor variable. The table reports the difference between the estimated mean of the predictive coefficient and the true mean. In each set of samples there is a long-history asset calibrated to the S&P 500 and a short-history asset calibrated to the EAFE or sub-index. The long-history asset has 124 years of data; the short-history asset has 30 years of data. Predictive coefficients are reported in Table 2 and standard deviations and correlations of errors in Table 3. The predictor variable has an autocorrelation coefficient of 0.89. "Short" denotes standard GMM; "AM" denotes the adjusted-moment method; "OI" denotes the over-identified method.

	Short	AM	OI
US	0.120	0.028	0.015
EAFE	0.083	0.008	-0.003
Asia-Pacific	0.063	0.004	-0.005
Europe	0.098	0.011	-0.002
Europe without UK	0.119	0.023	0.008
Scandinavia	0.115	0.015	0.001